Prompt Injection is a Control Problem:
Topological Constraints on Latent Dynamics

Sylvain Cormier
Paraxiom Research

January 2026

Abstract

Prompt injection attacks are typically framed as semantic or alignment failures. We argue instead that they arise from unconstrained latent dynamics that permit adversarial state displacement. This work does not propose a defense mechanism but a reframing: prompt injection is a dynamical control problem arising from the absence of topological structure in latent space. We show that compact latent manifolds equipped with coherence-preserving dynamics suppress such attacks mechanically, without eliminating access to unseen regions necessary for creative behavior. We demonstrate this principle via a minimal toy model on $\mathbb{T}^2$, clarify the role of the coherence functional, distinguish our approach from superficially similar paradigms, and specify the limits of the current proposal. The contribution is to show that restoring topological coherence changes the class of possible attacks.

1 Introduction

Contemporary large language models operate in high-dimensional latent spaces with no enforced global structure. This architectural choice enables flexibility but also permits adversarial inputs to induce arbitrary state transitions. Prompt injection exploits this freedom: rather than persuading the model semantically, it displaces the latent trajectory into regions divorced from the intended task context [27, 12, 32].

Current defenses rely on input filtering, output monitoring, or alignment training [26, 4]. These approaches treat injection as a content problem. We propose an alternative framing: prompt injection succeeds because the latent space lacks topological constraints that would make such displacement dynamically unsustainable. This builds on the ERLHS framework for coherence-preserving dynamics [36] and extends recent work showing that topological constraints prevent hallucination [37] to the adversarial setting.

This paper develops the consequences of imposing such constraints. We show that compact topology and coherence-preserving dynamics:

1. mechanically suppress high-frequency adversarial perturbations,

2. preserve access to novel regions via continuous traversal,

3. transform injection from a single-shot exploit into a control problem with observable signatures.

2 Prompt Injection as Dynamical Vulnerability

2.1 The Unconstrained Latent Space

Let $z_t \in \mathbb{R}^d$ denote the latent state of a language model at step t. In standard transformer architectures, the dynamics $z_{t+1} = f(z_t, x_t)$ are learned without explicit constraints on:

- the geometry of the reachable set,
- the curvature of valid trajectories,
- the existence of forbidden or high-cost regions.

This means that for any two states $z, z' \in \mathbb{R}^d$, there exists (in principle) an input sequence that transitions between them, regardless of their semantic relationship.

2.2 Adversarial State Displacement

Prompt injection exploits this lack of structure. An adversarial input x^* is one that induces:

$$\|z_{t+1} - z_t\| > \delta \quad \text{or} \quad d_{\mathcal{M}}(z_{t+1}, \mathcal{T}) > \epsilon$$

where $\mathcal{T}$ is the task-relevant manifold and $d_{\mathcal{M}}$ is a manifold distance. The attack succeeds not by being semantically plausible but by exploiting the absence of a mechanism preventing large geodesic jumps.

Recent mechanistic interpretability work supports this view: successful jailbreaks correlate with activation patterns that diverge sharply from in-distribution behavior [35, 2].

3 Unbounded Drift vs. Structured Exploration

A natural objection is that constraining latent dynamics may suppress creativity by restricting access to novel states. This concern rests on a false equivalence.

3.1 Two Modes of Novelty

Definition 1 (Unbounded Drift)**.** *A system exhibits unbounded drift if for any $\epsilon > 0$ and any state z, there exists an input sequence of length n such that $\|z_n - z_0\| > M$ for arbitrarily large M, with no constraint on path curvature.*

Definition 2 (Structured Exploration)**.** *A system exhibits structured exploration if novel states are reachable only via continuous paths satisfying local curvature bounds: $\|z_{t+1} - z_t\| \leq \delta$ and $|\kappa(z_t)| \leq \kappa_{\max}$ for all t.*

Results from dynamical systems theory show that structured exploration on compact manifolds can access arbitrarily many distinct states while maintaining bounded variation [3, 17]. Creativity—understood as access to previously unvisited configurations—does not require unbounded displacement.

3.2 Compact Manifolds as Latent Spaces

Consider a compact manifold $\mathcal{M}$ (e.g., a torus $\mathbb{T}^d$ or a low-genus surface). Such spaces are:

- **Finite in volume** but **unbounded in traversal**: geodesics can extend indefinitely without leaving the space.
- **Globally navigable**: any two points are connected by a continuous path.
- **Structure-preserving**: the metric enforces neighborhood relations.

Unseen regions are not eliminated; they become reachable only through valid paths.

4 The Coherence Functional

4.1 What $H(z)$ Measures

The coherence functional $H : \mathcal{M} \to \mathbb{R}$ does not encode semantic correctness. It measures *internal consistency of representation dynamics.* The key constraint is bounded variation:

$$|H(z_{t+1}) - H(z_t)| \leq \epsilon$$

This limits trajectory curvature without fixing the trajectory itself. Novel states are admissible provided they preserve local consistency.

4.2 Concrete Instantiations

$H(z)$ may be realized as:

1. **Lyapunov functional**: A function $V(z)$ satisfying $\dot{V} \leq 0$ along trajectories, guaranteeing asymptotic stability [22, 18].
2. **Hamiltonian energy**: In Hamiltonian Neural Networks [13], the learned Hamiltonian $\mathcal{H}(q, p)$ is conserved along trajectories:
$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial q}\dot{q} + \frac{\partial \mathcal{H}}{\partial p}\dot{p} = 0$$
3. **Spectral smoothness**: Given a graph Laplacian L over latent states, penalize high-frequency components [29]:
$$H(z) = z^\top L z = \sum_i \lambda_i |\hat{z}_i|^2$$
where λ_i are eigenvalues and $\hat{z}_i$ are spectral coefficients.
4. **Fisher-Rao regularization**: Constrain the statistical divergence between consecutive states [1]:
$$H(z_t, z_{t+1}) = \sqrt{g_{ij}(z_t)\Delta z^i \Delta z^j}$$
where g_{ij} is the Fisher information metric.

The framework is agnostic to the specific choice; the requirement is that H be computable and that its variation be bounded.

5 Spectral Properties and Perturbation Decay

5.1 The Spectral Gap Condition

Let Δ be the Laplace-Beltrami operator on $\mathcal{M}$, with eigenvalues $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots$. The spectral gap $\gamma = \lambda_1$ controls the decay rate of non-constant perturbations [8, 14].

For a perturbation ϕ orthogonal to the constant mode:

$$\|\phi(t)\| \leq e^{-\gamma t}\|\phi(0)\|$$

High-frequency adversarial inputs correspond to components with large λ_i, which decay faster.

5.2 Spectral Gap as Design Property

We do not claim that current LLMs possess a spectral gap. Rather:

- **Explicit construction**: Product manifolds (e.g., $\mathbb{T}^d = S^1 \times \cdots \times S^1$) have known spectral gaps [7].
- **Regularization**: Spectral normalization [24] and manifold regularization [5] can encourage gap-like properties in learned representations.

Spectral decay is a *design target*, not an empirical observation about existing systems.

6 A Minimal Toy Model

To ground these ideas, we present a minimal system exhibiting the claimed dynamics.

6.1 Setup

Consider latent space $\mathcal{M} = \mathbb{T}^2$, the 2-torus, parameterized by angles $(\theta, \phi) \in [0, 2\pi)^2$. Define Hamiltonian dynamics:

$$H(\theta, \phi, p_\theta, p_\phi) = \frac{1}{2}(p_\theta^2 + p_\phi^2) + V(\theta, \phi)$$

where V is a smooth potential. Hamilton's equations give:

$$\dot{\theta} = p_\theta, \quad \dot{\phi} = p_\phi, \quad \dot{p}_\theta = -\frac{\partial V}{\partial \theta}, \quad \dot{p}_\phi = -\frac{\partial V}{\partial \phi}$$

Energy H is conserved: $\frac{\mathrm{d}H}{\mathrm{d}t} = 0$.

6.2 Injection as Impulse Perturbation

Model prompt injection as an impulsive momentum kick:

$$p_\theta \mapsto p_\theta + \Delta p, \quad |\Delta p| \gg 1$$

This instantaneously increases energy: $H \mapsto H + \frac{1}{2}(\Delta p)^2 + O(\Delta p)$.

6.3 Response Under Topological Constraints

Without coherence constraint: The system continues at elevated energy. The injection succeeds permanently.

With coherence constraint $|H(t) - H_0| \leq \epsilon$: The perturbation violates the constraint. Two responses are possible:

1. **Rejection**: The transition is disallowed; the system remains at H_0.
2. **Dissipation**: Adding controlled damping $\dot{p} = -\gamma p$ causes excess energy to decay:

$$H(t) - H_0 \sim e^{-2\gamma t}$$

In either case, the injection fails to produce sustained state displacement.

6.4 Sustained Low-Energy Attack

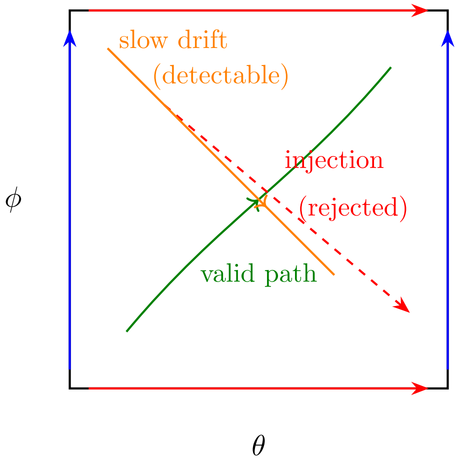
An adversary could attempt a slow, coherence-preserving drift:

$$p_\theta(t) = p_0 + \epsilon \cdot t, \quad \epsilon \ll 1$$

This respects $|\dot{H}| \leq \epsilon'$ and is not rejected. However:

- The attack requires $O(1/\epsilon)$ steps to achieve significant displacement.
- The trajectory is smooth and detectable via coherence trace analysis (Section 7).
- On $\mathbb{T}^2$, bounded drift eventually returns near the origin (Poincaré recurrence).

Conclusion: Impulse injection fails mechanically. Sustained injection becomes a control problem with observable cost.



Fundamental domain of $\mathbb{T}^2$

Figure 1: Dynamics on $\mathbb{T}^2$ under coherence constraints. Valid trajectories (green) follow smooth paths with bounded curvature. Impulsive injection attempts (red, dashed) violate energy bounds and are rejected. Slow drift attacks (orange) respect local constraints but require many steps and leave detectable traces.

7 Coherence Trace Analysis

To operationalize detection of residual attacks, we define:

Definition 3 (Coherence Trace)**.** *A coherence trace is a time-ordered sequence* $\tau = \{(z_t, \kappa_t, \Delta H_t, A_t)\}_{t=1}^{T}$ *where:*

- $z_t \in \mathcal{M}$ *is the latent state,*
- $\kappa_t = \|z_{t+1} - 2z_t + z_{t-1}\|$ *is local curvature,*
- $\Delta H_t = H(z_{t+1}) - H(z_t)$ *is energy variation,*
- $A_t = \mathbf{1}[z_{t+1} \notin \mathcal{N}_\delta(z_t)]$ *is an adjacency violation flag.*

Anomalous trajectories exhibit one or more of:

1. **Curvature spikes**: $\kappa_t > \kappa_{\max}$ indicates attempted discontinuous transition.
2. **Sustained drift**: $\sum_{t=1}^{T} \Delta H_t > \epsilon T$ with consistent sign indicates directed displacement.
3. **Frequency leakage**: High-frequency spectral components $|\hat{z}_k|$ for $k > k_{\max}$ indicate attempted injection that partially bypassed rejection.
4. **Adjacency violations**: $\sum_t A_t > 0$ indicates topological discontinuity.

These signals are computable from the latent trajectory without access to input semantics.

8 Relation to Existing Methods

Several existing paradigms involve constraints on generation. Our proposal differs in level of operation and the nature of guarantees.

8.1 Constrained Decoding

Methods such as FUDGE [33], NeuroLogic [21], and GeLaTo [34] impose constraints at the token level during autoregressive generation. These operate on:

- discrete token sequences,
- local next-token distributions,
- output-level properties (lexical, syntactic, semantic).

Distinction: We propose constraints on continuous latent trajectories, not discrete outputs. Token-level constraints cannot prevent latent displacement that occurs before decoding.

8.2 Energy-Based Models

EBMs [20, 11] define an energy function over outputs and sample via MCMC or Langevin dynamics. Recent work applies EBMs to text [10, 28].

Distinction: EBMs define energy over *outputs* or *input-output pairs*. We define coherence over *latent state transitions*. An EBM may assign low energy to an adversarially-reached output; our framework would flag the trajectory that reached it.

8.3 Diffusion Models

Diffusion models [16, 30] generate via iterative denoising along a stochastic path. They implicitly define a manifold of reachable outputs.

Distinction: Diffusion paths are stochastic and lack explicit topological invariants. There is no conserved quantity analogous to H. Our framework requires deterministic or controlled-stochastic dynamics with verifiable conservation laws.

8.4 Summary of Distinctions

Method	**Level**	**Constraint Type**	**Trajectory Verifiable**
Constrained decoding	Token	Output properties	No
Energy-based models	Output	Energy minimization	No
Diffusion models	Output	Stochastic path	No
This work	Latent state	Topological + dynamical	Yes

Table 1: Comparison of constraint paradigms. Our approach operates on latent trajectories with verifiable invariants.

9 Creativity and Discontinuity

9.1 The Discontinuity Objection

Some creative outputs appear to involve discontinuous conceptual jumps. Does requiring continuous latent paths preclude such creativity?

9.2 Continuity in Higher Dimensions

Evidence from cognitive science [19, 31] and representation learning [6, 15] suggests that apparent discontinuities often correspond to continuous paths in sufficiently high-dimensional or appropriately structured latent spaces.

A "creative leap" from concept A to concept B may traverse:

- intermediate abstractions,
- shared structural features,
- analogical bridges in representation space.

9.3 Scoped Claim

We do not claim that all creativity is continuous. We claim:

> *Creativity sufficient for language generation does not require unbounded latent displacement.*

This is empirically testable: if a topologically constrained model produces outputs judged as creative by human evaluators, the claim is supported.

10 Adversarial Adaptation

10.1 Limits of the Proposal

We do not claim that topological constraints eliminate all attacks. Adversaries may attempt:

- low-frequency, coherence-preserving drift,
- exploitation of regions where H is poorly defined,
- attacks on the coherence functional itself.

10.2 What the Framework Provides

Topological constraints enforce:

1. **Path dependence**: Attacks must follow valid trajectories. There are no shortcuts.
2. **Increased attack cost**: Achieving displacement D requires $O(D/\epsilon)$ steps under curvature bound ϵ.
3. **Observable signatures**: Coherence traces reveal sustained directional drift, anomalous curvature, or frequency leakage.

Prompt injection becomes a control problem rather than a single-shot exploit. The attacker must solve an optimal control problem under state and path constraints; the defender gains a detection surface.

11 Architectural Scope

11.1 What This Is Not

This proposal is not:

- a patch to existing transformers,
- a training-time intervention,
- a prompt engineering technique.

11.2 What This Is

This proposal describes a different design space characterized by:

- **Explicit latent topology**: The manifold $\mathcal{M}$ is specified, not emergent.
- **Coherence-aware dynamics**: Transitions are governed by H-preserving rules.
- **Trajectory-level verification**: Security properties are checked on paths, not outputs.

11.3 Bridging Work

Connecting this framework to existing architectures requires:

1. methods for projecting transformer representations onto compact manifolds [25, 23],
2. integration of Hamiltonian or Lagrangian layers [13, 9],
3. development of efficient coherence monitoring.

This remains future work.

12 Conclusion

Prompt injection is not defeated by rules, alignment, or output filtering, but by restoring structure to latent dynamics. We have argued that:

1. Injection succeeds due to unconstrained latent topology, not semantic failure.
2. Compact manifolds with coherence-preserving dynamics mechanically suppress high-frequency adversarial perturbations.
3. Creative exploration is preserved: unseen regions remain reachable via continuous paths.
4. Residual attacks become control problems with observable signatures.

The contribution is not a defense mechanism but a reframing. By treating prompt injection as a dynamical control problem arising from topological permissiveness, we identify a class of architectural constraints that change the nature of possible attacks.

Key claim (scoped): *Unseen regions remain reachable, but only through coherent paths. Restoring topology to latent dynamics transforms prompt injection from exploit to control problem.*

References

[1] S. Amari. Information Geometry and Its Applications. Springer, 2016.

[2] A. Arditi et al. Refusal in language models is mediated by a single direction. *arXiv:2406.11717*, 2024.

[3] V. Arnold. Mathematical Methods of Classical Mechanics. Springer, 2nd edition, 1989.

[4] Y. Bai et al. Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*, 2022.

[5] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[6] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE TPAMI*, 35(8):1798–1828, 2013.

[7] M. Berger. Eigenvalues of the Laplacian. In *Global Analysis*, Proc. Symp. Pure Math., 1971.

[8] F. Chung. Spectral Graph Theory. AMS, 1997.

[9] M. Cranmer et al. Lagrangian neural networks. *ICLR Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[10] Y. Deng et al. Residual energy-based models for text generation. *ICLR*, 2020.

[11] Y. Du and I. Mordatch. Implicit generation and generalization in energy-based models. *NeurIPS*, 2019.

[12] K. Greshake et al. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv:2302.12173*, 2023.

[13] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. *NeurIPS*, 2019.

[14] A. Grigor'yan. Heat Kernel and Analysis on Manifolds. AMS, 2009.

[15] I. Higgins et al. Towards a definition of disentangled representations. *arXiv:1812.02230*, 2018.

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[17] A. Katok and B. Hasselblatt. Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, 1995.

[18] H. Khalil. Nonlinear Systems. Prentice Hall, 3rd edition, 2002.

[19] B. Lake, T. Ullman, J. Tenenbaum, and S. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[20] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

[21] X. Lu et al. NeuroLogic decoding: (Un)supervised neural text generation with predicate logic constraints. *NAACL*, 2021.

[22] A. Lyapunov. The general problem of the stability of motion. PhD thesis, Kharkov, 1892. (Reprinted in *Int. J. Control*, 1992.)

[23] E. Mathieu et al. Continuous hierarchical representations with Poincaré variational auto-encoders. *NeurIPS*, 2019.

[24] T. Miyato et al. Spectral normalization for generative adversarial networks. *ICLR*, 2018.

[25] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 2017.

[26] L. Ouyang et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[27] F. Perez and I. Ribas. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. *EMNLP*, 2022.

[28] L. Qin et al. Cold decoding: Energy-based constrained text generation with Langevin dynamics. *NeurIPS*, 2022.

[29] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[30] Y. Song et al. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021.

[31] J. Tenenbaum, C. Kemp, T. Griffiths, and N. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

[32] A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How does LLM safety training fail? *NeurIPS*, 2023.

[33] K. Yang and D. Klein. FUDGE: Controlled text generation with future discriminators. *NAACL*, 2021.

[34] H. Zhang et al. GeLaTo: Generative latent textual optimization for text generation. *ACL*, 2023.

[35] A. Zou et al. Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*, 2023.

[36] S. Cormier. ERLHS: A Hamiltonian framework for coherence-preserving machine intelligence. *Zenodo*, 2025. DOI: 10.5281/zenodo.17928909.

[37] S. Cormier. Topological constraints for coherent language models: Why geometry prevents hallucination. *Paraxiom Research*, 2026.